IDENTIFICATION OF MULTIPLE INFLUENTIAL OBSERVATIONS IN WEIBULL REGRESSION

SITI NABILAH SYUHADA BINTI ABDULLAH

A project report in partial
fulfillment of the requirement for the award of the degree of
Master of Science (Industrial Statistics)

Faculty of Science, Technology and Human Development
Universiti Tun Hussein Onn Malaysia

JULY 2015

ABSTRACT

This study focuses on using the Cook's Distance and the Generalised DFFITS in identifying multiple influential observations on a linear regression set up. Non linear distributions are commonly used to model the survival analysis. One such distribution is the Weibull distribution which is used in this study. With the lifetime generated, the performance of Cook's Distance and the Generalised DFFITS in multiple influential observations under percentage of contaminate of 15% and 10%; and various sample sizes of n = 40,50 and 100 are measured. The objectives of the study are to identify multiple influential observations in Weibull Regression and to compare diagnostic method of Cook-type measure and Generalised DFFITS. A simulation study was conducted and the Log-linear form of Weibull distribution was used to generate the lifetime data. The method was also applied to a real data set, namely the patients diagnosed and treated for the human immunodeficiency virus (HIV) between January 1989 until November 1995. On the whole, it is concluded that the Generalised DFFITS diagnostic is the best method of detecting multiple influential observation in a Weibull Regression Model.

ABSTRAK

Kajian ini memberi tumpuan kepada penggunaan Jarak Cook dan DFFITS Am dalam mengenal pasti sekelompok pemerhatian berpengaruh pada regresi sejajar. Pengagihan bukan sejajar biasanya digunakan untuk model analisis survival. Salah satu daripadanya adalah taburan Weibull yang digunakan dalam kajian ini. Dengan jangka hayat yang dijana, prestasi Jarak Cook dan DFFITS Am dalam pemerhatian berpengaruh berkelompok di bawah peratusan pencemaran sebanyak 15% dan 10%; dan pelbagai saiz sampel n = 40,50 dan 100 telah diukur. Objektif kajian ini adalah untuk mengenal pasti pemerhatian berpengaruh berkelompok di dalam Regresi Weibull dan membandingkan kaedah diagnostik Jarak Cook dan DFFITS Am. Satu kajian simulasi telah dijalankan dan bentuk Log linear sebaran Weibull telah digunakan untuk menjana data hayat. Kaedah ini juga digunakan untuk satu set data sebenar, iaitu pesakit yang menghidap virus human immunodeficiency (HIV) antara Januari 1989 hingga November 1995. Pada keseluruhannya diagnostik DFFITS Am adalah kaedah terbaik untuk mengesan pemerhatian berpengaruh berkelompok dalam Model. Regresi Weibull.

CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background

In statistics, nonlinear regression is a form of regression analysis in which observational data are modeled by a function which is a nonlinear combination of the model parameter and depends on one or more independent variables. The data are fitted by a method of successive approximations. To build a correct model, the nature of the relationship and the data itself must be studied in depth. The relationship between the variables can be accessed through graphical means such as a scatter plot.

Another important step in regression analysis is to conduct a robustness study to detect influential or extreme observations that can cause important distortions on the results of the analysis. An observation is influential if it is individually or together with several other observations, has a demonstrably larger impact on the calculated values of various estimates than the case for most of the other observation (Belsley, 1980). The detection of influential observations, are of great importance and its development has been considered for both parametric model (Hall, Rogers and Pregibon, 1982; Weissfeld and Schneider, 1988) and semi parametric proportional hazard model (Reidand Crepeau, 1985).

Presence of outliers in the data may indicate a number of problems such as sample peculiarity or a mistake in data entry. It could also indicate observations having a low probability of occurrence but cannot be statistically shown to originate from a different distribution than the rest of the data in which case they cannot be eliminated from the data. Outliers may occur due to a shift in the mean or variability of the process and not to technical problems in which case it should not be removed

from the system. These types of outliers are called influential observations. Numerous approaches have been proposed in the literature with a view to detect influential or outlying observations that can seriously affect parameter estimates. Studies of case deletion have started by Cook (1977). Important reviews on the main approaches to detect influential observations are considered in Cook and Weisberg (1982) and Chatterjee and Hadi (1988). However, the diagnostic most commonly discussed pare for the identification of a single influential observation which becomes ineffective when swamping occur.

The Weibull distributions have been widely used in the analysis of survival data especially in medical and engineering application. This family of distribution is suitable in situations where the risk function is constant or monotone. This study focuses on the Weibull model. As considered by Mudhokar *et al.* (1995), it can be used in adjusting survival data with bathtub-type risk functions. Cancho *et al.* (1999) conducted a Bayesian study for exponentiated-Weibull regression models and Bolfarine and Cancho (2001) considered an exponentiated-Weibull survival model with a survival fraction. In this study, the local influence approach (Cook 1986) in detecting multiple influential observations in Weibull regression models was used. The relevance of the approach is illustrated with a simulated data set. This study uses a simulation data generated using R and a set of real time data. The data are assed for presence of multiple influential observations using Cook's Distance and the Generalised DFFITS.

1.2 Problem Statement

This study focuses on using the Cook's Distance and the Generalised DFFITS in identifying multiple influential observations on a linear regression set up. Non linear distributions are commonly used to model the survival analysis. One such distribution is the Weibull distribution which is used in this study. With the lifetime generated and perturbing the covariate values, the performance of Cook's Distance and the Generalised DFFITS in multiple influential observations under several percentage of contaminate and sample size.

1.3 Objectives of the Study

The objectives of the study are as follows:

i. To identify multiple influential observations in Weibull Regression with actual lifetime data
ii. To compare the performance of diagnostic method of Cook-type measure and Generalised DFFITS in detecting multiple influential observations.

1.4 Significance of study

The Weibull distribution is suitable in situations where the risk function is constant or monotone and have been widely used in the analysis of survival data especially in medical and engineering application. The study would provide a method that could best be applied in order to identify multiple influential observations in non linear regression with Weibull distribution

1.5 Scope of study

This research focuses on identifying multiple influential observations in Weibull distribution. In the nonlinear regression set up, the tools for detection of influential observations can include outlier and high leverage point observation. Thus, identification of multiple outliers is important in statistical analysis. A data containing outliers could lead to misleading results and misinterpretation.

1.6 Organization of Study

This study is organized as to be comprised of five chapters. Literature review related to this study will be presented in Chapter 2, followed by a detailed explanation of the research methodology employed in this study in Chapter 3. Chapter 4 presents the findings and results from the analysis. Finally, the study concludes and provides suitable suggestions in Chapter 5.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

Nonlinear regression has regressed over the years becoming one of the most preferred methods of analysis in analyzing relationship among quantitative and qualitative responses to make predictions and draw out important inferences.

There are several criteria to be considered before a model can be decided for any data set. When regression is considered to be applied, some general questions are posed in the process of defining the particular form of model most appropriate for the data. For instance, whether a linear or nonlinear approach would be more suitable and whether or not the data are affected by any unusual points. The relationship among data series also needs to be study in depth to uncover any collinear relationship among them exist since it may affect the parameter estimate and leave the model to be poorly specified.

Diagnostic techniques were developed over the years to handle such model-fitting problems and asses the quality of regression estimates. The nonlinear regression statistics are computed and used as in linear regression and the approximation towards linear causes the introduction of bias in to the model. Therefore, more caution than usual is required in interpreting statistics driven from a nonlinear regression. Although techniques for regression diagnostics have been originally developed for the conventional linear regression models, recent studies have explored and extend these diagnostics to nonlinear regression.

In regression diagnostic, influential observations are commonly related to the terms outliers and high leverage points. An influential observation is one which

either individually or together with several other observations demonstrate large impacts on calculated values or estimates than in the case for most of the other observations, (Belsley, 1980). Accurate identification of is important in avoiding misspecification of model. (Lawrance, 1991)

2.2 Weibull Distribution

In probability theory and statistics, the Weibull distribution is a continuous probability distribution. It is named after Waloddi Weibull, who described it in detail in 1951, although it was first identified by Fréchet (1927) and first applied by Rosin & Rammler (1933) to describe a particle size distribution.

The Weibull shape parameterγ is also known as the Weibull slope. This is because the value of γ is equal to the slope of the line in a probability plot. Different values of the shape parameter can have marked effects on the behavior of the distribution. The parameter γ is a pure number and it is dimensionless (Bolfarine, 2001).

Another characteristic of the distribution where the value of γ has a distinct effect is the failure rate. If the quantity is a "time-to-failure", *(t)* the Weibull distribution gives a distribution for which the failure rate is proportional to a power of time. This is one of the most important aspects of the effect of γ on the Weibull distribution (Escobar, 1992). A value of $\gamma < 1$ indicates that the failure rate decreases over time. This happens if there is significant "infant mortality", or defective items failing early and the failure rate decreasing over time as the defective items are weeded out of the population. A value of $\gamma = 1$ indicates that the failure rate is constant over time. This might suggest random external events are causing mortality, or failure. Finally, a value of $\gamma > 1$ indicates that the failure rate increases with time. This happens if there is an "aging" process, or parts that are more likely to fail as time go on (Bolfarine, 2001).

The form of the density function of the Weibull distribution changes drastically with the value of γ. For $0 < \gamma < 1$, the density function tends to ∞ as x approaches zero from above and is strictly decreasing. For $\gamma = 1$, the density function tends to $1/\lambda$ as x approaches zero from above and is strictly decreasing For $\gamma > 1$, the density function tends to zero as x approaches zero from above,

increases until its mode and decreases after it. It is interesting to note that the density function has infinite negative slope at $x = 0$ if $0 < \gamma < 1$, infinite positive slope at $x = 0$ if $1 < \gamma < 2$ and null slope at $x = 0$ if $\gamma > 2$. For $\gamma = 2$ the density has a finite positive slope at $x = 0$. As γ goes to infinity, the Weibull distribution converges to a Dirac delta distribution centered at $x = \lambda$. Moreover, the skewness and coefficient of variation depend only on the shape parameter.

A change in the scale parameter, λ has the same effect on the distribution as a change of the abscissa scale. Increasing the value of λ while holding γ constant has the effect of stretching out the p.d.f. Since the area under a p.d.f curve is a constant value of one, the "peak" of the p.d.f curve will also decrease with the increase of λ.

If λ is increased, while γ is kept constant, the distribution gets stretched out to the right and its height decreases, while maintaining its shape and location. If λ is decreased, while γ is kept constant, the distribution gets pushed in towards the left and its height increases has the same unit as T, such as hours, miles, cycles, actuations, etc. (Bolfarine, 2001).

Proportional hazards models are a class of survival models in statistics. Survival models relate the time that passes before some event occurs to one or more covariates that may be associated with that quantity of time. In a proportional hazards model, the unique effect of a unit increase in a covariate is multiplicative with respect to the hazard rate. Other types of survival models such as accelerated failure time models do not exhibit proportional hazards. The proportional hazards model, proposed by Cox (1972), has been used primarily in medical testing analysis, to model the effect of secondary variables on survival. It is more like an acceleration model than a specific life distribution model, and its strength lies in its ability to model and test many inferences about survival without making any specific assumptions about the form of the life distribution model. The Cox model may be specialized if a reason exists to assume that the baseline hazard follows a particular form. In this case, the baseline hazard is replaced by a Weibull hazard function which gives the Weibull proportional hazards model. Incidentally, using the Weibull baseline hazard is the only circumstance under which the model satisfies both the proportional hazards, and accelerated failure time models.

2.3 Diagnostic of Influential Observation

A general framework to detect influence of observations was proposed by Cook (1986) and has often been applied with regression models. The method basically indicates how sensitive the analysis is when small perturbations are made to the data or the model. For instance, under the normal error, Lawrance (1988) investigated local influence applications in linear models with a response transformation parameter, Beckman *et al.* (1987) presented influence studies in mixed effects analysis of variance, Tsai and Wu (1992) considered first-order autoregressive models with non-constant variances, and Paula (1993) used local influence methods with linear regression models when there are inequality constraints on the parameters. Moving away from normal models, Petit and Bin Daud (1989) investigated local influence with proportional hazard regression models, Escobar and Meeker (1992) adapted local influence methods to regression analysis with censoring, and O'Hara *et al.* (1992) and Kim (1995) applied local influence methods with multivariate regression. More recently, Galea *et al.* (1997) and Liu (2000) used local influence with elliptical linear regression models; Kwan and Fung (1998) applied the methodology to factor analysis and Gu and Fung (1998) discussed local influence in canonical correlation analysis. An interesting discussion and comparison with other influence measures is considered in Fung and Kwan (1997). An important extension of the method to assess the local influence of observations on the predictions from the fitted model was proposed by Thomas and Cook (1990).

To determine the Cook's Distance in Proportional Hazard (Weibull Regression), the linear regression and logistic regression cases are studied. Though the types of statistic used in linear and logistic regression differ quite a bit from proportional hazard regression, the essentials are similar in all three. In linear and logistic regression, leverage, which is used to determine an outlier by measuring values of covariates for any observation; is calculated as the distance between value of covariates to the overall mean of the covariates. In the proportional hazard regression however, the Leverage are not easily defined and does not possess the same properties due to the possibility of subject appearing in multiple risk sets hence being present in multiple terms in the partial likelihood.

To determine influence in the proportional hazard regression, analogous Cook's Distance *(CD$_i$)* statistic is used. The *(CD$_i$)* approximates the change in value of the estimated coefficients if a subject is deleted from the data. (CD_i) may also be used to provide a single overall summary statistic on the affect of a single subject on the estimator of the entire coefficient. The overall measure for Cook's Distance in proportional hazard model is

$$(\mathrm{CD}_i) = \left(\hat{\beta}_i - \hat{\beta}_{-i}\right)^{\mathrm{T}} \hat{\Sigma}_{\hat{\beta}}^{-1} \left(\hat{\beta}_i - \hat{\beta}_{-i}\right) \tag{2.1}$$

Where $\hat{\beta}_i$ denotes the partial likelihood estimator of coefficient estimated using the entire sample size and $\hat{\beta}_{-i}$ denoted the value of the estimator if the i^{th} subject is removed.

The diagnostic tools discussed so far are designed for the identification of a single influential observation and are ineffective when masking and/or swamping occur. Therefore, we need detection techniques that are free from these problems. A group-deleted version of the residuals and weights are used to effectively diagnose the identification of multiple influential observations in Weibull regression was introduced by Nurunnabi in 2010 .The identification of multiple outliers starts by constructing set of D cases through graphical display. Then, the set D would be accessed using both the Cook-type measure (Cook, 1982) and the generalized DFFITS (Imon, 2005).

Once the measure for set D has been calculated, it will be compared to Mean Absolute Deviation (MAD) which is considered as the cutoff point. Should the measure yield a larger value compared to the MAD, it is considered as an influential observation. The GDFFITS is used (Imon, 2005) for the entire data set. It is defined as:

$$GDFFITS_i = \begin{cases} \dfrac{\hat{y}_i^{(-D)} - \hat{y}_i^{(-D-i)}}{\hat{\sigma}^{(-D-i)}\sqrt{h_{ii}^{(-D+i)}}}, i \in R \\ \dfrac{\hat{y}_i^{(-D+i)} - \hat{y}_i^{(-D)}}{\hat{\sigma}^{(-D)}\sqrt{h_{ii}^{(-D+i)}}}, i \in D \end{cases} \tag{2.2}$$

Where $h_{ii}^{(-D)} = x_i^T (X_R{}^T X_R)^{-1} x_i$ and $h_{ii}^{(-D+i)} = x_i^T (X_R{}^T X_R + x_i^T x_i)^{-1} x_i = \dfrac{h_{ii}^{(-D)}}{1 + h_{ii}^{(-D)}}$

For identification of multiple outliers, the GDFFITS are expressed by Standardised Pearson Residual (GSPR) and the generalized weights (GWs) (Nurunnabi, 2010) defined as:

$$GDFFITS_i = r_{si}^{(-D)} \sqrt{h_{ii}^{*(-D)}} \tag{2.3}$$

Where $h_{ii}^{*(-D)} \begin{cases} \dfrac{h_{ii}^{(-D)}}{1-h_{ii}^{(-D)}}, i \in R \\ \dfrac{h_{ii}^{(-D)}}{1+h_{ii}^{(-D)}}, i \in D \end{cases}$ and $r_{si}^{(-D)} = \begin{cases} \dfrac{\hat{y}_i^{(-D)} - \hat{y}_i^{(-D-i)}}{\hat{\sigma}^{(-D)}\sqrt{1-h_{ii}^{(-D)}}}, i \in R \\ \dfrac{\hat{y}_i^{(-D+i)} - \hat{y}_i^{(-D)}}{\hat{\sigma}^{(-D)}\sqrt{1+h_{ii}^{(-D)}}}, i \in D \end{cases}$

The cases having $|GDFFITS_i| \geq c\sqrt{\dfrac{k}{n-d}}$ are considered as influential observation. A suitable constant for c is between 2 and 3 (Imon, Identifying Multiple Influential Observation in Linear Regression, 2005).

CHAPTER 3

METHODOLOGY

3.1 Introduction

The entire procedure involved is explained thoroughly in this chapter. The generation of data following a Weibull distribution and the simulation study of identifying multiple influential observation are discussed in the following pages of this chapter.

3.2 Parameterization of Simulation Study

The parameterization of Weibull distribution with hazard function is employed for this study

$$h(t) = \frac{\gamma}{\lambda^{\gamma}}(t)^{\gamma-1} \qquad (3.1)$$

For a vector of $\underline{x} : (x_1, x_2, \ldots, x_p)$, two approaches of Weibull regression model can be used:

- Proportional Hazard Model
- Log linear model

3.2.1 Proportional Hazard Model

The Weibull proportional hazards model is the Cox model specialized to assume the baseline hazard following a Weibull form. The hazard function for the i^{th} individual is;

$$h_i(t) = e^{(\beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip})} \frac{\gamma}{\lambda^\gamma}(t)^{\gamma-1} \tag{3.2}$$

Suppose the values $x_1, x_2, \ldots, x_p$ of exploratory variables $X_1, X_2, \ldots, X_p$ are recorded for each of n individuals. Under the proportional hazards model, consider the survival time $T>0$, sometimes censored, and suppose that a vector of basic covariates $x_i^T = (x_1, x_2, \ldots, x_{ip})$ is available on each individual taken at or before time 0. The relationship between T and covariate x can be modeled and determined whereby x becomes the predictive of subsequent survival times. Thus to include the covariate vector x_i of the i^{th} individual, the hazard given x_i can be expressed as;

$$h\left(\frac{t}{x_i}\right) = \frac{\gamma}{\lambda^\gamma}(t)^{\gamma-1} e^{(\sum \beta_j x_{ij})} = \frac{\gamma}{\left[\lambda\left(e^{-\sum \beta_j x_{ij}}\right)^{1/\gamma}\right]^\gamma}(t)^{\gamma-1} \tag{3.3}$$

where the survival of the i^{th} individual follows a Weibull distribution with scale parameter $\lambda\left(e^{-x'\beta}\right)^{1/\gamma}$ and shape parameter γ. The hazard is finally defined as:

$$\begin{aligned} h(t \mid \underline{x}) &= \frac{\gamma}{\lambda^\gamma}(t)^{\gamma-1} e^{x'\beta} \\ &= \gamma\left\{\frac{1}{\lambda}\left(e^{x'\beta}\right)^{\frac{1}{\gamma}}\right\}(t)^{\gamma-1} \\ &= \gamma\left(\frac{1}{\tilde{\lambda}}\right)^\gamma t^{\gamma-1} \end{aligned} \tag{3.4}$$

Where $\left(\frac{1}{\tilde{\lambda}}\right) = \frac{1}{\lambda}\left(e^{x'\beta}\right)^{\frac{1}{\gamma}}$ or $\tilde{\lambda} = \lambda\left(e^{x'\beta}\right)^{-\frac{1}{\gamma}}$

3.2.2 Log Linear Model

Consider a log linear model for the random variable T_i associated with the survival time of i^{th} individual in a survival study that can be written as

$$Y_i = \beta_0^* + \sum \beta_j^* x_{ij} + \sigma\varepsilon_i \tag{3.5}$$

where $\beta_1^*, \beta_2^*, \ldots, \beta_p^*$ are unkown coefficients of p explanatory variable $X_1, X_2, \ldots, X_p$, and $\beta_0^*$is intercept and, σ is scale parameter. The two models are related through $\beta_0^* = \log\lambda; \beta^* = -\frac{1}{\gamma}\beta, \sigma = \frac{1}{\gamma}$; and $\lambda = e^{\beta_0^*}; \beta = -\gamma\beta^*; \gamma = 1/\sigma$. ε_i on the other hand is a random variable used to model the departures of the value of $\log T_i$ from the linear part of the model and is assumed to have a particular probability function, specifically in this case, the Gumbell Distribution. The probability density function is given by $f(\varepsilon) = e^{\varepsilon - e^{\varepsilon}} for -\infty < \varepsilon < \infty$. If we define $\xi = e^{\varepsilon}$ then the probability density function of ξ is $-e^{\varepsilon}$ which is exponential with unit mean. (See Collet, 1994)

A data of $Y = \beta_0^* + x'\beta^* + \sigma z$ following a Weibull Distribution $W \sim (\gamma, \tilde{\lambda})$ is generated whereby $\gamma = \frac{1}{\sigma}$ and $\tilde{\lambda} = \lambda\left(e^{x'\beta}\right)^{-\frac{1}{\gamma}} = e^{\beta_0^*} e^{x'\beta} = e^{\beta_0^* + x'\beta}$

3.3 Generate Data from Weibull Distribution

Steps to generate a sample with complete observations:

Step 1. Generate a random sample n with $X \sim N(\mu, \sigma)$ with predefined values of μ,and σ

Step 2. Estimate $\eta = \beta_0^* + x'\beta^*$

Step 3. Estimate $\tilde{\lambda} = e^{\beta_0^* + x'\beta}$

Step 4. Estimate $\gamma = 1/\sigma$

Step 5. Generate T~Weibull $\left(W \sim (\gamma, \tilde{\lambda})\right)$ with replication *(1,n)* where 1 represents uncensored data

3.4 Simulation Study

The lifetime data is generated with sample size of $n = 40, 50,$ and 100 with two level of influential percentage at 15% and 10% contamination procedure. The methods of detection are applied using the generated data. Index plot and character plot of explanatory and response variables could give ideas on influential observations but are highly unreliable on repressors of high dimension and depends purely on the judgment of the researcher and hence is subjected to bias. Influential observations are potential outliers or high leverage points or both and can be detected through some leverage measures such as the *Median Absolute Deviation (MAD)* employed by this study.

3.4.1 Cook-type Measure

Steps involved in identifying multiple influential observations using the Generalized Cook-type Measure:

Step 1. Estimate β with the full data

Step 2. Delete the i^{th} observation from data and estimate the new β

This is a measure of distance between $\hat{\beta}_{(i)}$ and $\hat{\beta}$. A high value indicates that the estimated beta is far from the actual value and hence indicates that the i^{th} unit has a substantial effect on the full sample.

Step 3. Repeat for $i=1,2,3,...n$

Step 4. Calculate CD_i

Step 5. Compare the CD_i of each unit in Set D with the robust cut off point defined as

$$Median(CDi)+3MAD(CDi)$$

Where $MAD(CD_i) = \dfrac{Median\{|CD_i - Median(CD_i)|\}}{0.6745}$

When CD_i has a value higher than MAD, it is concluded that the observations are influential. If the condition is not satisfied, the cases are put back into the estimation subset.

3.4.2 Generalized DFFITS

Steps involved in identifying multiple influential observations using the Generalised DFFITS Measure:

Step 1. For simplicity, use the Set D identified earlier

Step 2. Compute GDFITS for entire data, where

$$GDFFITS_i = \begin{cases} \dfrac{\hat{y}_i^{(-D)} - \hat{y}_i^{(-D-i)}}{\hat{\sigma}^{(-D-i)}\sqrt{h_{ii}^{(-D+i)}}}, i \in R \\ \dfrac{\hat{y}_i^{(-D+i)} - \hat{y}_i^{(-D)}}{\hat{\sigma}^{(-D)}\sqrt{h_{ii}^{(-D+i)}}}, i \in D \end{cases} \tag{3.6}$$

For identification of influential observations, using c=3 (Nurunnabi, 2010), it must follow the rule

$$|GDFFITS_i| \geq c\sqrt{\frac{k}{n-d}} \tag{3.7}$$

If the condition is not satisfied, the cases are put back into the R subset.

3.5 Determining Suitable Method of detecting Multiple Influential Observation in Weibull Regression

By considering several scenarios, the simulation study could identify the most adequate method to identify multiple influential observations in Weibull regression. In this study, the data are divided into two parts; part 1 with influential observation and part 2 consist of influential observation. For example, for n=40 with influential percentage of 10%, the last 4 observations are left to be influential (part 2) while the first 36 are not (part 1).

Parameter values are predefined as $\beta_0^* = 0, \beta_1^* = 2, \beta_2^* = 2, \beta_3^* = 2, \sigma = \frac{2}{3}$ and the scenarios considered are $n = 40, 50, 100$ percentage of influential at 10% and 15% and contamination process. Contamination procedure is as follows:

1) Change in β_0^*; new $\beta_0^* = \beta_0^* + 1$
2) Change in only β_1^*; $new\beta_1^* = \beta_1^* + 1$
3) Change in $\beta_1^*, \beta_2^*, \beta_3^*$; new $\beta_j^* = \beta_j^* + 1, j = 1, 2, 3$
4) Change in σ; new $\sigma = \sigma + 1$
5) Change in all parameters $\theta^* = \beta_0^*, \beta_1^*, \beta_2^*, \beta_3^*$ and σ;

   new $\beta_j^* = \beta_j^* + 1, j = 1, 2, 3$; new $\theta^* = \theta^* + 1$

Using the previous example of n=40 with 10% influential, the first through thirty six observation would follow the initial parameter estimate of $\beta_0^* = 0, \beta_1^* = 2, \beta_2^* = 2, \beta_3^* = 2, \sigma = \frac{2}{3}$. However the last 4 observation will be contaminated by following the procedure explained.

The simulation procedure was conducted for a total of 1000 times. For each 100 run, the number of times each formula was able to detect the influential observations correctly was recorded and denoted as Yinf. The number of times each formula was not able to detect the correct influential observations was also recorded and denoted as Ninf. Then, the detection rate and false detection rate are calculated. The formula are defined as:

$$\text{Influential Detection Rate: } \frac{\text{Y inf}_{Cumulative}}{1000 \times \text{influential}\% \times n}$$

$$\text{False Detection Rate: } \frac{N\text{inf}_{Cumulative}}{1000 \times \text{influential}\% \times 30}$$

The best method is chosen based on its ability to increase influential rate detection and the worst method is the one with small false rate detection.

CHAPTER 4

DATA ANALYSIS AND RESULTS

4.1 Introduction

This section will provide the key findings in this study in relation with identifying multiple influential observations in a Weibul Regression set up. Through simulation study under several scenarios of sample size and percentage of contamination, the performance of both Cook-type measure and generalized DFFITS method are analyzed. Both diagnostic methods are applied on a set of secondary data in which the results will be presented in this chapter.

4.2 Simulation study

The lifetime data is generated with sample size of $n = 40, 50,$ and 100 with two level of influential percentage at *15%* and *10%* contamination procedure. The methods of detection are applied using the generated data. The simulation procedure was conducted 1000 times. The percentages of detection of multiple influential observations are defined as Influential Detection Rate (Yinf) and False Detection Rate (Ninf). For each 100 run, the number of Yinf and Ninf detectied is recorded and finally the influential and false detection rate is calculated. The Cook-type measure and the generalized DFFITS performance are compared under the several scenarios taken into account in the simulation study.

Four various ways of comparing were identified:

- Average Influential Observation based on influential percentage with specific sample size.
- Average influential Observation based on overall influential percentage (10% and 15%)
- Average Influential Observation based on sample size (n=40,50,100)
- Overall Average Influential Observation

The best method is chosen based on its ability to increase influential rate detection and the worst method is the one with small false rate detection.

4.2.1 Average Influential Observation based on influential percentage with specific sample size

Table 4.1 show the average influential observations based on influential percentage with specific sample size. Note that the method with a higher influential rate is best and the method with the smaller false detection rate detection is worst. Hence, from the table, it can be seen that at sample size 40, for both 10% and 15% influential, Cook-type measure is a better method compared to the generalized DFFITS. This is due to Cook-type measure having higher influential detection rate. For sample size 50 and 100 however, the result is the complete opposite where DFFITS yield a higher influential detection rate. It can therefore be concluded that Cook-type measure works best for smaller sample size while DFFITS works better with larger sample sizes in detecting multiple influential observations in Weibul Regression model.

Table 4.1 Average Influential Observation based on influential percentage with specific sample size

Influential Percentage	Sample Size, n	Method			
		DFFITS		Cook-type measure	
		Yinf. Rate	Ninf.Rate	Yinf. Rate	Ninf.Rate
10%	40	0.056282	**0.37193**	**0.091704**	0.35498
	50	**0.230321**	0.197054	0.21038	**0.220702**
	100	**0.392904**	0.053142	0.35266	**0.072952**
15%	40	0.050133	**0.387113**	**0.100634**	0.390981
	50	**0.246015**	0.252252	0.235677	**0.267627**
	100	**0.393691**	0.098113	0.363732	**0.100152**

4.2.2 Average influential Observation based on overall influential percentage (10% and 15%)

A closer value of influential detection rate to 100% indicates a better performance of the multiple influential observations detection and a smaller value of false detection rate indicates the poor performance of the method. Based on Table 4.2 and figure 4.1, the result shows that at influential percentage 10% DFFITS is the best model for detecting multiple influential observations in Weibull Regression Model and, based on the same table and figure 4.2, the result shows that Cook-type Measure is the best method for detecting multiple influential observations in Weibull Regression model at 15% influential.

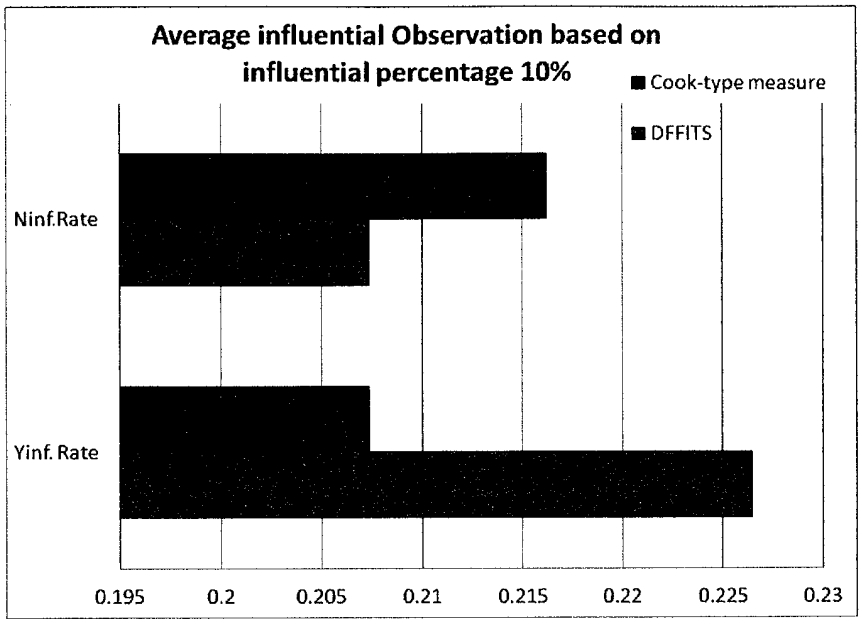

Figure 4.1 Average Influential Observations at 10% Influential

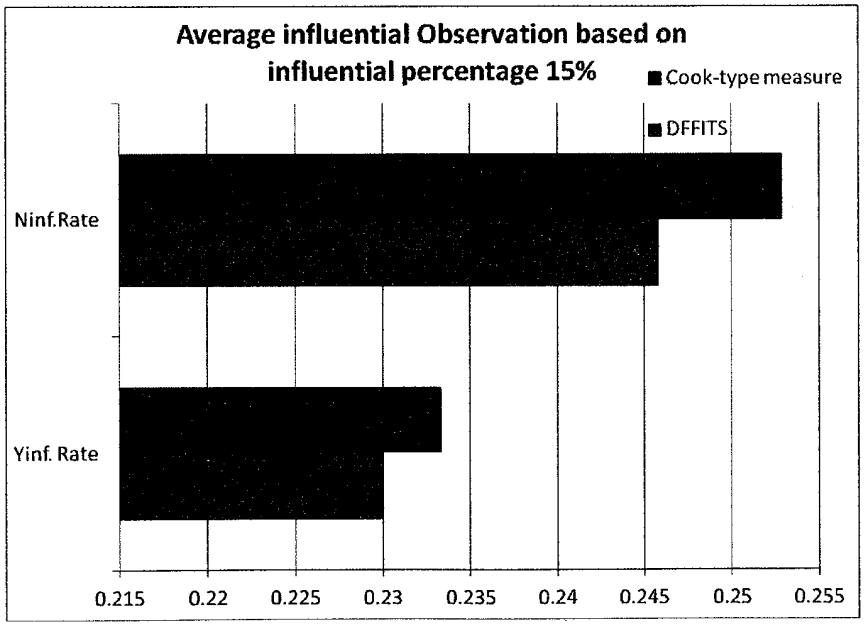

Figure 4.2 Average Influential Observations at 15% Influential

Table 4.2 Average influential Observation based on overall influential percentage

Influential Percentage	Method			
	DFFITS		Cook-type measure	
	Yinf. Rate	Ninf.Rate	Yinf. Rate	Ninf.Rate
10%	**0.226502**	0.207375	0.218248	**0.216211**
15%	0.229946	**0.245826**	**0.233348**	0.25292

4.2.3 Average Influential Observation based on sample size

Based on figure 4.3, the results show that as the sample size increase, the influential detection rate increase. Meanwhile, the false detection rate decreases as the sample size decrease. Note that the best method for detecting multiple influential observations is decided based on the higher influential detection rate. Based on table 4.3 and figure 4.4, the best method for detecting multiple influential observations in Weibull Regression model for sample size 40 is the Cook-type measure. For sample size 50 and 60, based on table 4.3, figure 4.5 and 4.6, the best method for detecting multiple influential observations in Weibul Regression model is the generalized DFFITS.

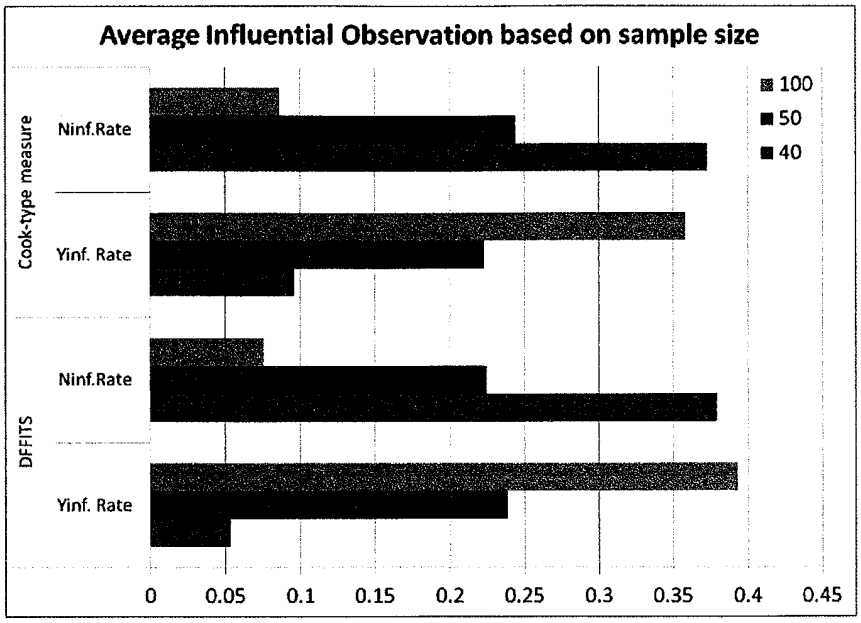

Figure 4.3 Average Influential Observations based on Sample Size

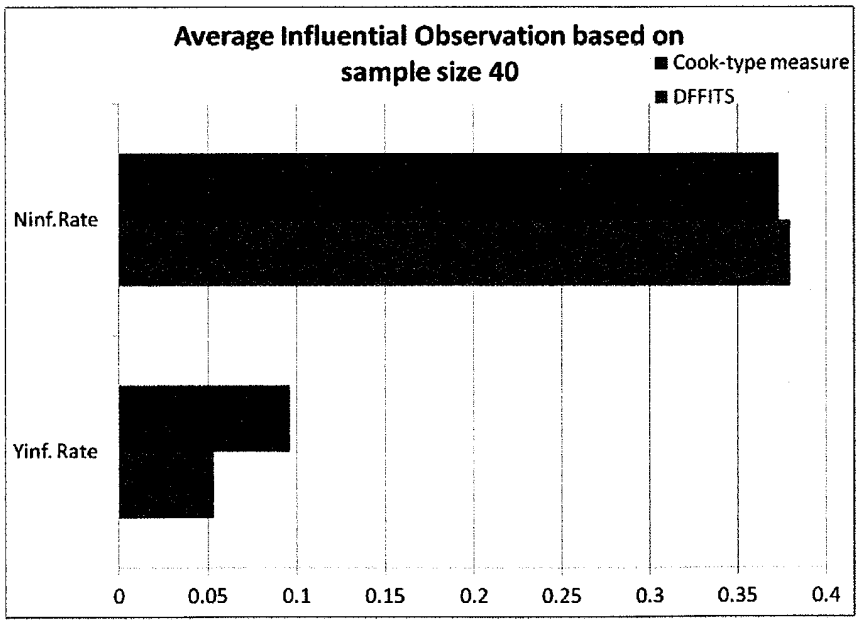

Figure 4.4 Average Influential Observations at n = 40

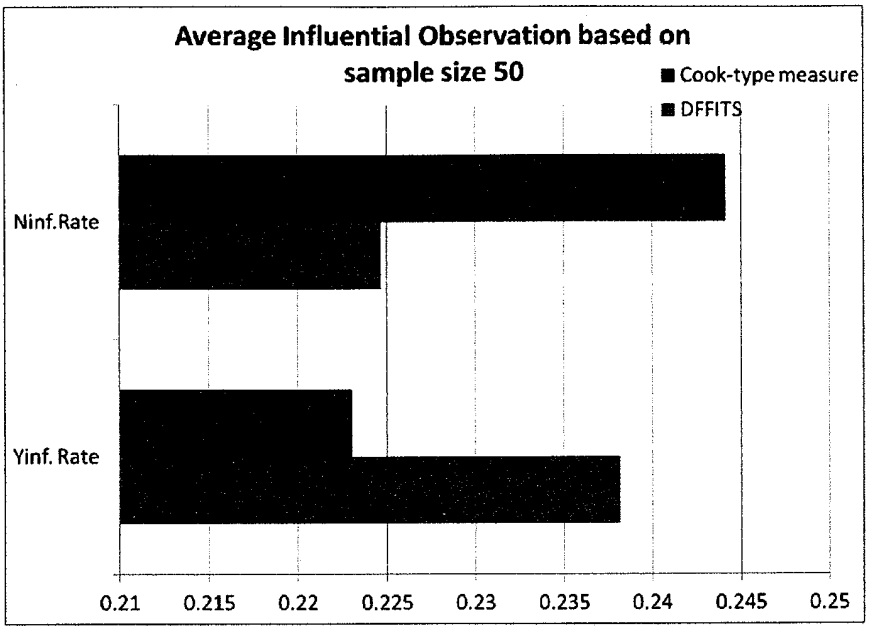

Figure 4.5 Average Influential Observations at n = 50

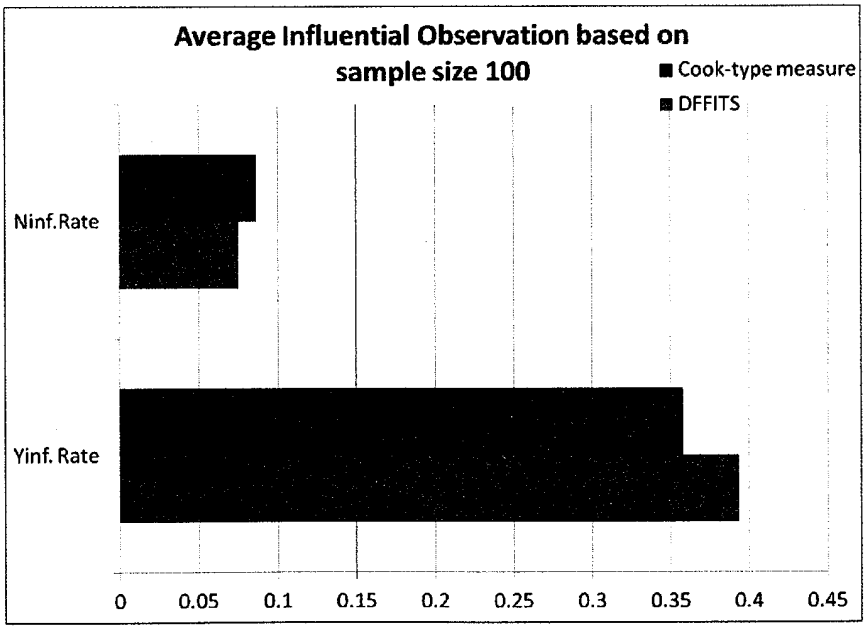

Figure 4.6 Average Influential Observations at n = 100

Table 4.3 Average Influential Observation based on sample size

Sample Size, n	Method			
	DFFITS		Cook-type measure	
	Yinf. Rate	Ninf.Rate	Yinf. Rate	Ninf.Rate
40	0.053207	**0.379521**	**0.096169**	0.372981
50	**0.238168**	0.224653	0.223029	**0.244164**
100	**0.393298**	0.075628	0.358196	**0.086552**

4.2.4 Overall Average Influential Observation

Table 4.4 and figure 4.7 shows the overall average influential observations result to compare both methods in detecting multiple influential observations. Since the DFFITS has a higher influential detection rate, it is considered as the best method to detect multiple influential observations in a Weibull regression Model.

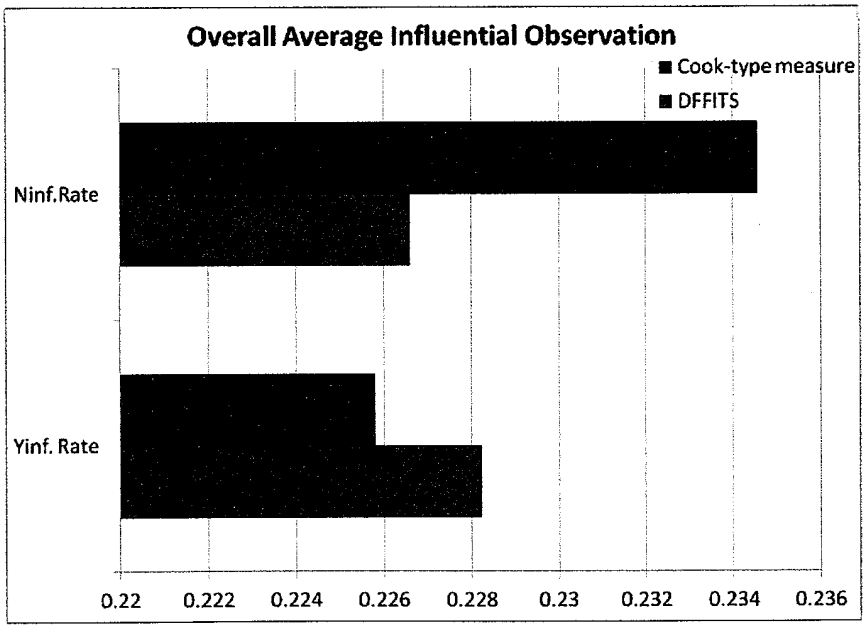

Figure 4.7 Average Influential Observations (Overall)

REFFERENCES

A.J. Lawrance, *Local and deletion influence* in *Directions in Robust Statistics and Diagnostics*, Part I,W. Stahel and S.Weisberg, eds., pp. 141–157, Springer, Berlin, 1991.

Beckman, R. J., Nachtsheim, C. J. and Cook, R. D. (1987). Diagnostics for mixed-model analysis of variance. *Technometrics*, 29, 413-426.

Bolfarine, H. and Cancho, V. (2001). Modelling the presence of immunes by using the exponentiated-Weibull model. *Journal of Applied Statistics*, 28, 659-671.

Cancho, V.; Bolfarine, H. and Achcar, J. A. (1999). A Bayesian analysis for the exponentiated-Weibull distribution. *Journal Applied Statistical Science*, 8, 227-242.

Chatterjee, S. and Hadi, A. S. (1988). *Sensitivity Analysis in Linear Regression.* New York: John Wiley.

Cook, R. (1982). Residuals and Influence in Regression. London: Chapman and Hall.

Cox, D. R. (1972). Regression models and life-tables (with discussion). J. R. Statist. Soc. B 34, 187-220

D.A. Belsley, E. Kuh, and R.E. Welsch, *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*,Wiley, NewYork, 1980.